Word Boundary Detection with Mel-Scale Frequency Bank in Noisy Environment

Gin-Der Wu and Chin-Teng Lin, *Senior Member, IEEE*

***Abstract*—This paper addresses the problem of automatic word boundary detection in the presence of noise. We first propose an *adaptive time-frequency* (ATF) parameter for extracting both the time and frequency features of noisy speech signals. The ATF parameter extends the TF parameter proposed by Junqua *et al.* from single band to multiband spectrum analysis, where the frequency bands help to make the distinction of speech and noise signals clear. The ATF parameter can extract useful frequency information by *adaptively* choosing proper bands of the mel-scale frequency bank. The ATF parameter increased the recognition rate by about 3% of a TF-based robust algorithm which has been shown to outperform several commonly used algorithms for word boundary detection in the presence of noise. The ATF parameter also reduced the recognition error rate due to endpoint detection to about 20%. Based on the ATF parameter, we further propose a new word boundary detection algorithm by using a neural fuzzy network (called SONFIN) for identifying islands of word signals in noisy environment. Due to the self-learning ability of SONFIN, the proposed algorithm avoids the need of empirically determining thresholds and ambiguous rules in normal word boundary detection algorithms. As compared to normal neural networks, the SONFIN can always find itself an economic network size in high learning speed. Our results also showed that the SONFIN's performance is not significantly affected by the size of training set. The ATF-based SONFIN achieved higher recognition rate than the TF-based robust algorithm by about 5%. It also reduced the recognition error rate due to endpoint detection to about 10%, compared to an average of approximately 30% obtained with the TF-based robust algorithm, and 50% obtained with the modified version of the Lamel *et al.* algorithm.**

***Index Terms*—Mel-scale frequency, multiband, neural fuzzy network, self-learning ability, spectrum analysis.**

I. Introduction

An important problem in speech processing is to detect the presence of speech in noisy environment, where the word boundary is hard to detect exactly. This problem is often referred to as the robust endpoint location problem. The inaccurate detection of word boundary will be harmful to recognition. The energy (in time domain), zero-crossing rate, and duration parameters have been usually used to find the boundary between the word signal and background noise [1]–[4]. It has been found that the energy and zero-crossing rate parameters are not sufficient to get reliable word boundaries in noisy environment, even if more complex decision strategies are used [5]. Especially, the zero-crossing rate is very sensitive to the additive noise. Up to date, several other parameters were proposed, such as linear prediction coefficient (LPC), linear prediction error energy [6], [7] and pitch information [8]. Although the LPC's are quiet successful in modeling vowels [9], they are not particular suitable for nasal sounds, fricatives, etc. The reliability of the LPC parameter depends on the noise environment. The pitch information can help to detect the word boundary, but it is not easy to extract the pitch period correctly in noisy environment. Four endpoint detection algorithms were compared in [5]: an energy-based algorithm with automatic threshold adjustment [3], [4], use of pitch information [8], a noise adaptive algorithm, and a voiced activation algorithm. These four algorithms are strongly dependent on the noise condition. The reliability of the parameters used by the four algorithms also depends on the noise condition.

In the connection, Junqua *et al.* [5] proposed the time-frequency (TF) parameter. They used the frequency energy in the fixed frequency band 250∼3500 Hz to enhance the time energy information. The TF parameter is the result obtained after smoothing the sum of the time energy and frequency energy. The frequency energy helps us to make the distinction between speech and noise. Based on the TF parameter, a robust algorithm was proposed in [5] to get more precise word boundary in noisy environment. This robust algorithm includes noise classification, a refinement procedure, and some preset thresholds. Although this algorithm outperforms several commonly used algorithms for word boundary detection in the presence of noise, it needs to empirically determine thresholds and ambiguous rules which are not easily determined by human. Some researchers used the neural network's learning ability to solve this problem. In [6], [7], [10], multilayer neural networks are used to classify the speech signal into voiced, unvoiced, and silence segments. In the neural network approach, the decision rules are in the form of input-output mapping, and can be learned by the training procedure (supervised leaning). However, the proper structure of the network (including numbers of hidden layers and nodes) is not easy to decide.

To develop a more robust word boundary detection algorithm and avoid the problems of the above approaches, this paper first proposes a modified TF parameter and then uses a neural fuzzy network to detect word boundary based on this parameter. Since the frequency energy (i.e., magnitudes of the spectrum) of different types of noise focus on different frequency bands, more accurate frequency information can be obtained by considering multiband analysis of noisy speech signals. With this

Manuscript received June 2, 1998; revised November 15, 1999. This work was supported by the National Science Council, R.O.C., under Grant NSC 89-2213-E-009-114. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wu Chou.

The authors are with the Department of Electrical and Control Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan, R.O.C. (e-mail: ctlin@fnn.cn.nctu.edu.tw).

Publisher Item Identifier S 1063-6676(00)06922-4.

motivation, we propose a new robust parameter, called *adaptive time-frequency* (ATF) parameter, which extends the TF parameter from single-band to multiband spectrum analysis based on the mel-scale frequency bank (20 bands). A procedure is proposed such that the ATF parameter can extract more informative frequency energy than the single-band approach to compensate the time-energy information by *adaptively* choosing proper frequency bands. The ATF parameter is the result obtained after smoothing the sum of the time energy and frequency energy. It makes the word signal more obvious than the TF parameter. According to our experiments, the new ATF parameter can increase by about 3% the recognition rate of the robust algorithm proposed by Junqua *et al*. [5], in which all the thresholds have been tuned exhaustively for our test environment. The ATF-based robust algorithm also reduced the recognition error rate due to endpoint detection to about 20%, compared to an average of approximately 30% obtained with the TF-based robust algorithm, and 50% obtained with the modified version of the Lamel *et al.* algorithm [3], [4].

Based on the ATF parameter, we further propose a new word boundary detection algorithm by using a neural fuzzy network for identifying islands of speech signals in noisy environment. The neural fuzzy network is called self-constructing neural fuzzy inference network (SONFIN) that we proposed previously in [11]. Due to the self-learning ability of SONFIN, the proposed algorithm avoids the need of empirically determining thresholds and ambiguous rules in normal word boundary detection algorithms. The SONFIN can always find itself an economic network size in high learning speed, and thus avoids the need of empirically determining the number of hidden layers and nodes in normal neural networks [12], [13]. Our experimental results showed that the proposed scheme achieved higher recognition rate by about 2% than the well-tuned ATF-based robust algorithm, or by about 5% than the original well-tuned TF-based robust algorithm. It also reduced the recognition error rate due to endpoint detection to about 10%, compared to about 20% obtained with the ATF-based robust algorithm. Our experiments also showed that the SONFIN's performance was not significantly affected by the size of training set.

This paper is organized as follows. The ATF parameter is derived in Section II, where the performance evaluation and comparisons of this new parameter are also done. In Section III, the structure and function of the SONFIN is briefly introduced, and then the SONFIN-based word boundary detection scheme is proposed. The performance evaluation and comparisons of the proposed scheme using the ATF parameter are performed extensively also in Section III. Finally, the conclusions of our work are summarized in Section IV.

II. Adaptive Time-Frequency (ATF) Parameter

In this section, we generalize the single-band analysis of the TF parameter to multiband analysis based on mel-scale frequency bank and propose a new *adaptive time-frequency* (ATF) parameter.

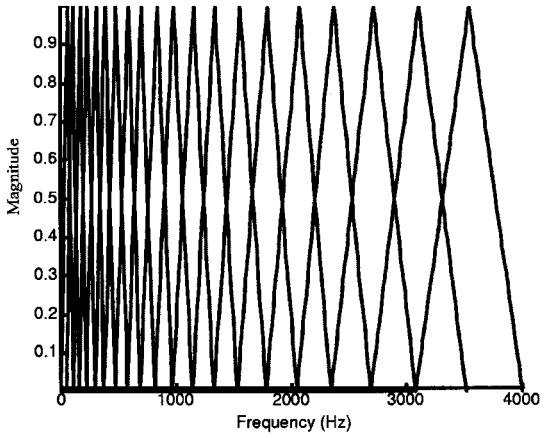


Fig. 1. Mel-scale filter-bank in which each filter has a triangular bandpass frequency response with bandwidth and spacing determined by a constant mel-frequency interval.

A. Auditory-Based Mel-Scale Filter Bank

There is a evidence from auditory psychophysics that the human ear perceives speech along a nonlinear scale in the frequency domain [14]. One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on a nonlinear, warped frequency scale, such as the mel scale. The relation between mel-scale frequency and frequency (Hz) is described by the following equation [15]:

$$mel = 2595 \log(1 + f/700) \tag{1}$$

where mel is the mel-frequency scale and f is in Hz. The filter bank is then designed according to the mel scale as shown in Fig. 1, where the filters of 20 bands are approximated by simulating 20 triangular band-pass filers, $f(i, k)$ $(1 \leq i \leq 20, 0 \leq k \leq 63)$, over a frequency range of 0$\sim$4000 Hz. Hence, each filter band has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval by (1). The value of the triangular function, $f(i, k)$, in the figure also represents the weighting factor of the frequency energy at the kth point of the ith band.

With the mel-scale frequency bank given in Fig. 1, we can now calculate the energy of each frequency band for each time frame of a speech signal. Consider a given time-domain noisy speech signal, $x_{\text{time}}(m, n)$, representing the magnitude of the nth point of the mth frame. We first find the spectrum, $x_{\text{freq}}(m, k)$, of this signal by discrete Fourier transform (128-point DFT)

$$x_{\text{freq}}(m, k) = \sum_{n=0}^{N-1} x_{\text{time}}(m, n) W_N^{kn}, \; 0 \leq k \leq N-1, \quad 0 \leq m \leq M-1, \tag{2}$$

$$W_N = \exp(-j2\pi/N) \tag{3}$$

where

$x_{\text{freq}}(m, k)$	magnitude of the kth point of the spectrum of the mth frame;
N	128 in our system;
M	number of frames of the speech signal for analysis.

We then multiply the spectrum $x_{\text{freq}}(m, k)$ by the weighting factors $f(i, k)$ on the mel-scale frequency bank and sum the products for all k to get the energy $x(m, i)$ of each frequency band i of the mth frame

$$x(m, i) = \sum_{k=0}^{N-1} |x_{\text{freq}}(m, k)| f(i, k), \; 0 \le m \le M-1, \; 1 \le i \le 20 \tag{4}$$

where

i filter band index;
k spectrum index;
m frame number;
M number of frames for analysis.

In order to remove some undesired impulse noise in (4), we further smooth it by using a three-point median filter to get $\hat{x}(m, i)$

$$\hat{x}(m, i) = \frac{x(m-1, i) + x(m, i) + x(m+1, i)}{3}. \tag{5}$$

Finally, the smoothed energy, $\hat{x}(m, i)$, is normalized by removing the frequency energy of background noise, Noise_freq, to get the energy of almost pure speech signal, $X(m, i)$, where the energy of background noise is estimated by averaging the frequency energy of the first five frames of the recording

$$\begin{aligned} X(m, i) &= \hat{x}(m, i) - \text{Noise\_freq} \\ &= \hat{x}(m, i) - \frac{\sum_{m=0}^{4} \hat{x}(m, i)}{5}. \end{aligned} \tag{6}$$

With the smoothed and normalized energy of the ith band of the mth frame, $X(m, i)$, we can calculate the total energy of the almost pure speech signal at the ith band as $E(i)$

$$E(i) = \sum_{m=0}^{M-1} |X(m, i)|. \tag{7}$$

Since our goal is to select some useful bands having the maximum word signal information, we need a parameter to stand for the amount of word signal information of each band. It is obvious that $E(i)$ in (7) is a good indicator for the amount of speech information, since the more the word signal information is covered by the noise, the smaller the $E(i)$ is. In other words, the larger the $E(i)$ is, the more word signal information the ith band has. Hence, we can extract useful frequency information for word boundary detection by adopting the bands having large $E(i)$.

Since the band with higher $E(i)$ contains more pure speech information, we shall sort the 20 mel-scale frequency bands according to their $E(i)$ values. This is also a preparatory task for the adaptively band-chosen method developed in the following subsection. Let S be the set of all $E(i)$

$$S = \{E(i) | i = 1, 2, 3, \cdots, 20\}. \tag{8}$$

The sorting is performed as follows:

$$\begin{aligned} P(1) &= \text{Max}\{S\} \\ P(2) &= \text{Max}\{S - \{P(1)\}\} \\ P(3) &= \text{Max}\{S - \{P(1), P(2)\}\} \\ &\vdots \\ P(20) &= \text{Max}\{S - \{P(1), P(2), \cdots, P(19)\}\} \end{aligned} \tag{9}$$

where $P(1)$ is the maximum total energy, and $P(20)$ is the minimum total energy. Let the band index corresponding to $P(i)$ be represented by $I(i)$ for $i = 1, 2, \cdots 20$. That is, $I(1)$ is the index of band having the maximum total energy $P(1)$, and $I(20)$ is that having the minimum total energy $P(20)$. We observed that not all of the 20 bands were helpful in making the distinction between word signal and background noise. The next problem is how to select the useful bands. We shall deal with this problem in the following subsection.

B. Adaptive Band Selection

Before we consider the adaptive choices of suitable bands for extracting useful frequency information, we first make some observations on the effect of additive noise on each frequency band. Obviously, larger background noise will add more noise component into each band, and thus reduce each $E(i)$. However, some bands are corrupted more seriously than the others. These seriously obscured bands have little word signal information left, and are not useful, if not harmful, for word boundary detection. In other words, the number of useful bands decreases as the energy of background noise increases. We denote the number of bands useful for producing reliable frequency energy as N_a. Large N_a should be used at high SNR. On the contrary, small N_a is used at low SNR because most bands are corrupted seriously by the additive noise. On the other experiments, we observed that even at the same noise energy level (SNR), the useful bands are different under different noise conditions. This is because different noise sources focus their energy on different frequency bands. Hence, there are two factors affecting the selection of useful bands, SNR, and noise characteristics. The effects of these two factors can be detected by the total frequency energy $E(i)$ in (7).

For illustration, the smoothed and normalized frequency energies of a clean speech signal, $X(m, i)$ in (6), for 20 bands ($i = 1, 2, \cdots, 20$) and 100 frames ($m = 0, 1, \cdots, 99$) are shown in Fig. 2(a). Specifically, the energies of the fifth and eighteenth bands, $X(m, 5)$ and $X(m, 18)$, are shown in Fig. 2(b). From the figure, we observe that the word signal is clear (in the sense of frequency energy) in both the fifth and eighteenth bands, whose maximum $X(m, i)$ values are about 40 and 30, respectively. If we consider the total frequency energy $E(i)$ in (7), both $E(5)$ and $E(18)$ are large with $E(5) \ge E(18) \ge 300$. Hence, both the fifth and eighteenth bands can help us to find the word signal part, and are recognized as useful frequency bands. We then add white noise (10 dB) to the same clean speech signal to see the effects of adding noise on each band. The corresponding $X(m, i)$ values of the 20 bands are shown in Fig. 3(a), and the new $X(m, 5)$ and $X(m, 18)$ values are given in Fig. 3(b). We observe that the additive noise reduces $X(m, 5)$ and $X(m, 18)$, and thus reduces $E(5)$ and $E(18)$, but we still have $E(5) \ge E(18)$. Hence, both the two bands are corrupted

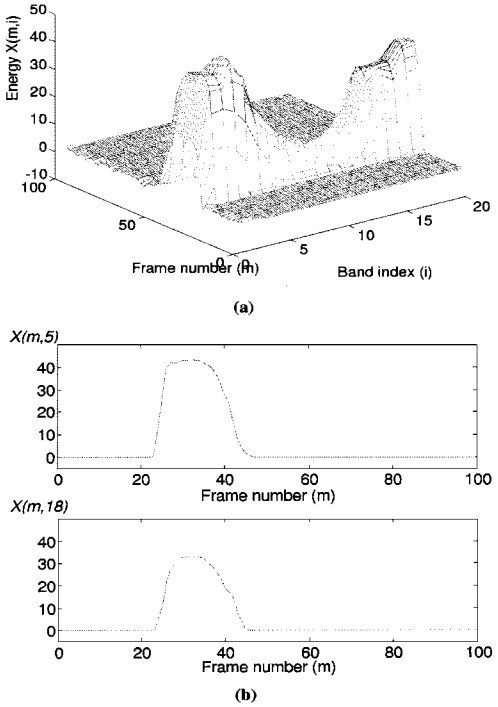


Fig. 2. Multiband spectrum analysis of a clean speech signal with length of 100 frames. (a) Smoothed and normalized frequency energies, $X(m, i)$, on 20 frequency bands. (b) Smoothed and normalized frequency energies, $X(m, 5)$ and $X(m, 18)$, on the fifth and eighteenth frequency bands.

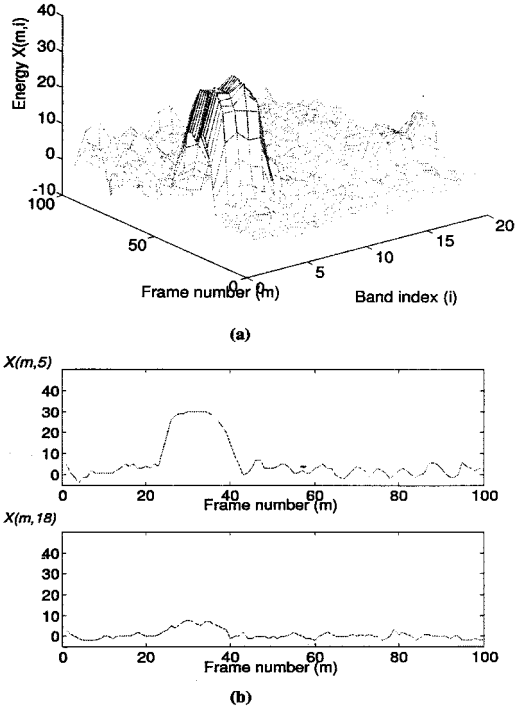


Fig. 3. Multiband spectrum analysis of the speech signal in Fig. 2 with additive white noise of 10 dB. (a) Smoothed and normalized frequency energies, $X(m, i)$, on 20 frequency bands. (b) Smoothed and normalized frequency energies, $X(m, 5)$ and $X(m, 18)$, on the fifth and eighteenth frequency bands.

by the additive noise. However, Fig. 3(b) shows that the eighteenth band is corrupted by the added noise more seriously than the fifth band ($E(5) \geq 300$ and $E(18) < 300$). The word signal is still clear in the fifth band whose maximum $X(m, 5)$ value is about 30, but the word signal is ambiguous in the eighteenth band whose maximum $X(m, 18)$ value falls below ten. As a result, we cannot extract helpful word signal information from the eighteenth band, and we will not use this band as a useful frequency band. On the other hand, the fifth band is still a useful frequency band in the added white-noise environment.

In order to provide some physical justification for that higher noise level will lead to a smaller value of N_a, the average number of bands whose total energy $E(i)$ is greater than 300 under different noise conditions and SNRs is shown in Table I (with a total of 600 utterances.) We can easily find that the number of useful bands decreases as the energy of the background noise increases.

Based on the above discussion and illustrations, we now propose a way to choose the number of useful bands adaptively for extracting helpful frequency information. More precisely, after ordering the band indexes according to their total frequency energy $(E(i))$ as in (9), we want to decide the number N_a such that the first N_a bands $(I(1), I(2), \cdots, I(N_a))$ can produce helpful frequency energy, $(P(1) = E(I(1)), P(2) = E(I(2)), \cdots, P(N_a) = E(I(N_a)))$. At first, we observed from our experiments that the first 18 bands (after ordering) could provide the maximum improvement for word boundary detection in clean environment. Little improvement was observed with the addition of the other two bands. We also observed that one or two bands only cannot give helpful frequency

TABLE I
EXPERIMENTAL STATISTICS ON THE AVERAGE NUMBER OF BANDS WHOSE $E(i) \geq 300$ UNDER DIFFERENT NOISE CONDITIONS AND SNRS

SNR	Noise			
	White	Babble	Cockpit	Factory
Clean	16.8	16.8	16.8	16.8
20dB	13.2	16.4	14.7	15.5
15dB	11.0	16.4	13.2	15.2
10dB	9.1	15.4	10.7	13.4
5dB	7.6	15.0	8.1	10.4
0dB	6.0	10.1	5.4	7.3

information in our test cases. Hence, we bound the N_a values between 3 and 18 for the noisiest and clean environments, respectively. Within the range (3, 18), N_a is tuned adaptively according to the strength of background noise; higher noise level should lead to smaller N_a value as observed in Table I. To obtain a reliable tuning rule for N_a, we first observed from our experiments that the average frequency energy of background noise, Noise_freq [see (6)], is 83 in clean environment, and is 93 at a low SNR value (5 dB). We set the corresponding numbers of useful bands to be 18 and 3 for these two extreme cases, respectively. For computation simplicity, we assume that the relation between N_a and Noise_freq is linear. With the above experimental observations and assumption, we sketch the relation diagram between N_a and Noise_freq in Fig. 4. From this figure, we can derive the tuning rule for N_a as follows:

$$\frac{N_a - 18}{\text{Noise\_freq} - 18} = \frac{18 - 3}{83 - 93}, \quad 3 \leq N_a \leq 18, \quad N_a \text{ is an integer} \tag{10}$$

which gives

$$N_a = -1.5 \times \text{Noise\_freq} + 142.5, \quad 3 \le N_a \le 18, \quad N_a \text{ is an integer.} \tag{11}$$

Rewriting the above result into a general form, we have

$$N_a = \lfloor A \times \text{Noise\_freq} + B \rfloor, \quad 3 \le N_a \le 18, \quad A = -1.5 \quad \text{and } B = 142.5 \tag{12}$$

where $\lfloor\ \rfloor$ is a function used to denote the rounding to nearest integer operation, and A and B are constants determining the slop and offset, respectively, of the linear relation between N_a and Noise_freq, where A is negative.

With the number of useful bands, N_a, decided by (12), we then sum the total energies of the first N_a bands (after ordering) in (9) to get the final frequency energy, $F(m)$, of frame m

$$F(m) = \sum_{i=1}^{Na} X(m, I(i)). \tag{13}$$

The proposed adaptive time-frequency (ATF) parameter of the mth frame is the result obtained after smoothing the sum of the frequency energy $F(m)$ in (13) and time energy $T(m)$

$$\text{ATF}(m) = \text{SMOOTHING}(T(m) + cF(m)) \tag{14}$$

where SMOOTHING is performed by a three-point median filter as in (5), constant c is a proper weighting factor, and the time energy $T(m)$ is given by smoothing and normalizing the logarithm of the root-mean-square (rms) energy of the time-domain speech signal

$$x_{\text{rms}}(m) = \log \sqrt{\frac{\sum_{n=0}^{L-1} x_{\text{time}}^2(m,n)}{L}}, \tag{15}$$

$$\hat{x}_{\text{rms}}(m) = \frac{x_{\text{rms}}(m-1) + x_{\text{rms}}(m) + x_{\text{rms}}(m+1)}{3} \tag{16}$$

$$T(m) = \hat{x}_{\text{rms}}(m) - \text{Noise\_time}, = \hat{x}_{\text{rms}}(m) - \frac{\sum_{m=0}^{4} \hat{x}_{\text{rms}}(m)}{5} \tag{17}$$

where L is the length of the frame, which is 120 (15 ms) in our system. The procedure to calculate the ATF parameter is illustrated in Fig. 5(a). The details of the block with label "Select N_a useful bands to produce frequency energy" of this figure is shown in Fig. 5(b).

C. Evaluation of the ATF Parameter

In this section, we shall test the performance of the proposed ATF parameter, and compare it to the original TF parameter. The tests are performed by using either ATF or TF parameter as input feature of a word boundary detection algorithm. The detected word signal is then sent into a speech recognizer. Since inaccurate detection of word boundary is harmful to recognition, the performance of the word boundary detection process, and thus the performance of the ATF parameter, is examined by the recognition rate of speech recognizer. In the following, we shall introduce the used word boundary detection algorithm, speech recognizer, test database, and the evaluation results.

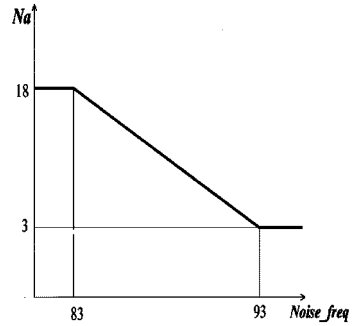
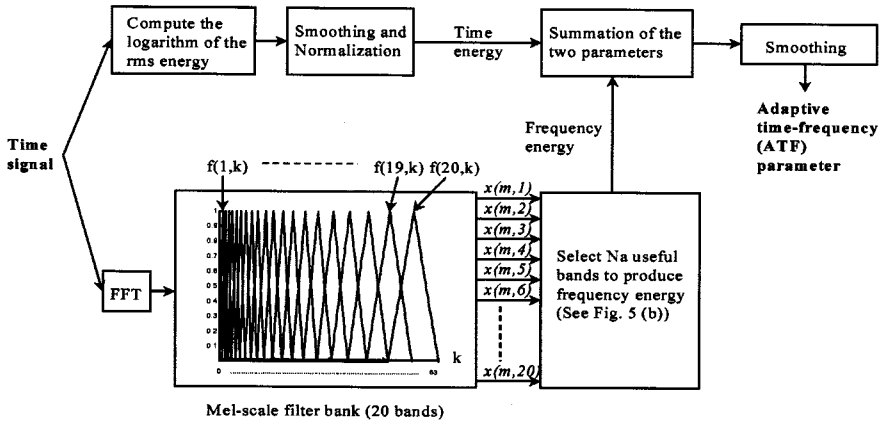


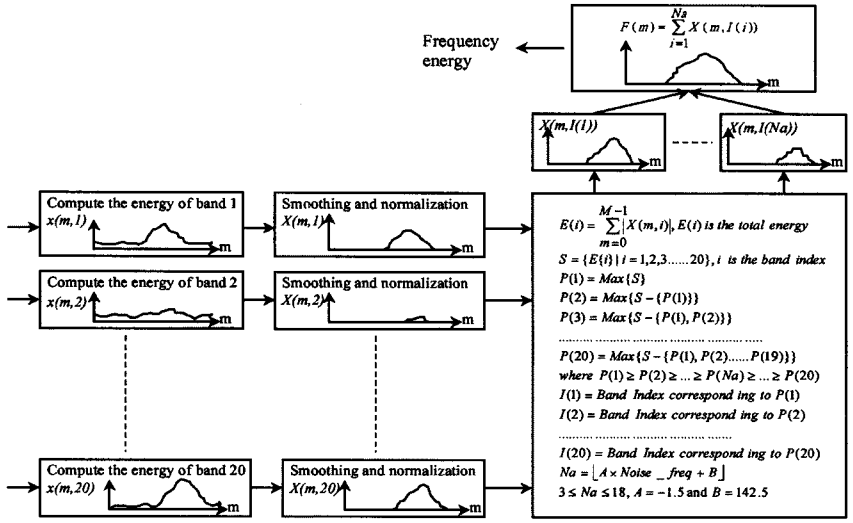
Fig. 4. Relation between N_a and *Noise_freq*.

1) Robust Word Boundary Detection Algorithm: The algorithm adopted for word boundary detection in this subsection is the robust algorithm proposed by Junqua *et al.* [5]. This robust algorithm used the TF parameter and was shown to outperform several commonly used algorithms for word boundary detection in the presence of noise. We shall feed this algorithm with the proposed ATF parameter instead of TF parameter later for performance comparisons. The robust algorithm first performs a noise classification procedure to determine noise level (high, medium, or low) and the noise category (high or low zero-crossing rate) by using ten frames of "relative" silence at the beginning of the recording, and computing an average of the logarithm of the rms energy and the zero-crossing rate on these frames. A set of empirically determined threshold values are used to perform the noise classification. After noise classification, the robust algorithm applies a noise adaptive procedure to determine the word boundary. It uses the TF parameter with some thresholds to find the islands of reliability boundary. Finally, the refinement procedure which also depends on the noise classification results is applied to the initial boundary. It tries to find the earliest boundary by subtracting adjustment value (typically 20 ms) from the beginning boundary to obtain new boundary (maximum up to 100 ms from the beginning island of reliability boundary.) Then, using some thresholds to determine the end of finding final beginning boundary. It tries to find the latest boundary by adding adjustment value (typically 50 ms) from ending boundary to obtain new boundary (maximum up to 150 ms from the ending island of reliability boundary). Then, using some thresholds to determine the end of finding final ending boundary. The thresholds in the refinement procedure include the logarithm of the time-domain rms energy and zero-crossing rate.

2) Speech Recognition System: The speech recognition system used in this paper for evaluating the performance of word boundary detection algorithms is a robust isolated word recognition system consisting of two parts, feature extractor and classifier. In the feature extractor, the modified two-dimensional cepstrum (modified TDC-MTDC) [16]–[19] is used as the speech feature. The MTDC can simultaneously represent several types of information contained in the speech waveform: static and dynamic features, as well as global and fine frequency structures. To represent an utterance, only some



(a)



(b)

Fig. 5. (a) Flowchart for computing the ATF parameter. (b) Adaptive band selection procedure in (a) for computing frequency energy.

MTDC coefficients need to be selected to form a feature vector instead of the sequence of feature vectors. The MTDC has the advantage of simple computation and is suitable for noisy speech recognition due to its choices of robust coefficients. In the classifier, a Gaussian clustering algorithm is used. The training was done on clean speech pronounced in a clean environment (without background noise.) In the training phase, each model is trained by a mixture of four Gaussian distribution density functions. We use a total of 1000 utterances for training. The details of the above isolated word recognition system can be found in [19].

3) Test Environment and Noise Speech Database: In the recognition procedure, the frame window used for obtaining the MTDC features is 30 ms in length, and is with 15 ms overlapping between two frames. In the word boundary detection procedure, the frame length is set to be 15 ms in order to get more accurate endpoint location. The sampling rate of our system is 8 KHz. The noise signals are taken from the noise database provided by the NATO Research Study Group on Speech Processing (RSG.10) NOISE-ROM-0 [20]. The database consists of 24 noise sources in order to offer as wide as possible variations in characteristics. Among these noise sources, we take four typical types of noise for speech contamination in our experiments. They are multitalker babble noise, cockpit noise, noise on the floor of car factory, and white noise. The original NOISE-ROM-0 data were sampled at 19.98 KHz and stored as 16-bit integers. In our experiments, they are prepared for use by downsampling to 8 KHz and applying attenuation on them. The attenuation was applied to enable the addition of noise without causing an overflow of the 16-bit integer range. The speech data used for our experiments are the set of isolated Mandarin digits. They are ten digits spoken by 10 speakers and each speaker

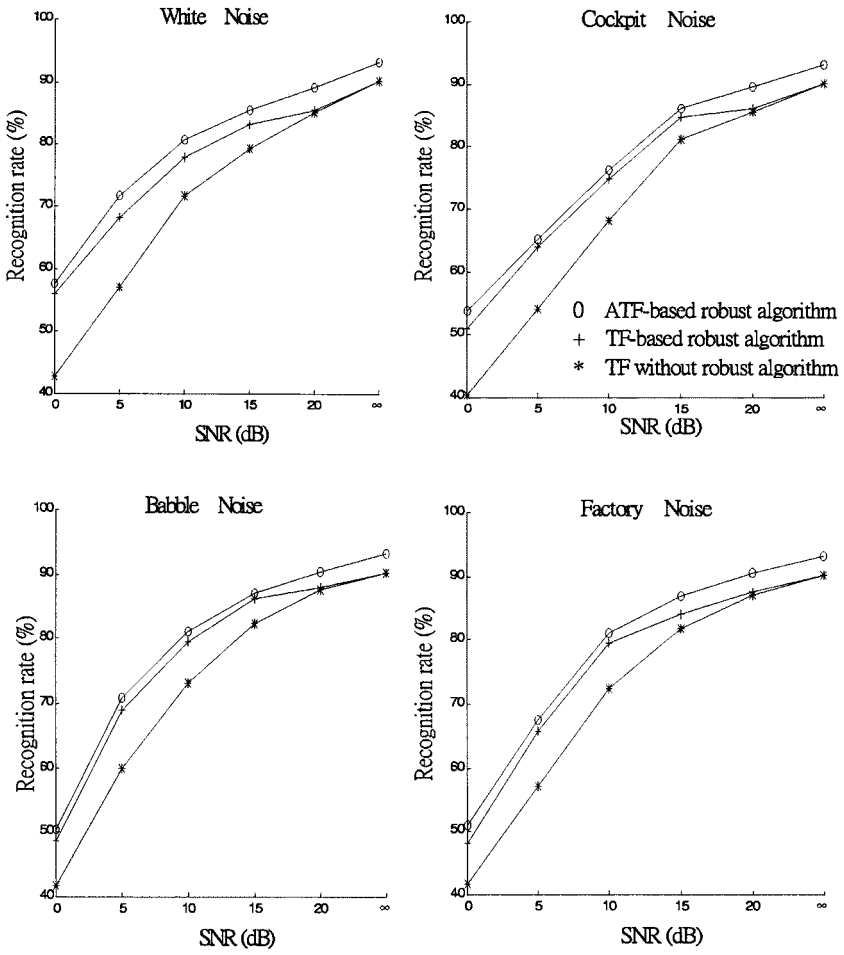


Fig. 6. Recognition rates of three word boundary detection algorithms (ATF-based robust algorithm, TF-based robust algorithm, and TF without robust algorithm) in an MTDC-based recognition system across six SNRs and four noise conditions.

pronounced 20 times of the ten digits. The recording sampling rate is 8KHz and stored as 16-bit integer. To set up the noisy speech database for testing, we added the prepared noisy signals to the recorded speech signals with different signal-to-noise-ratios (SNRs) including 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and ∞ dB. The duration of each utterance used for testing the performance of the word boundary detection algorithm is about one second (including silence.) A total of 600 utterances were used in our experiments.

4) Experimental Results: Three algorithms were used in our experiments; they are ATF-based robust algorithm, TF-based robust algorithm, and TF without robust algorithm. The TF without robust algorithm uses no noise classification and no refinement procedure. It is used only for a reference to see the performance of the robust algorithm. The recognition rates of these three algorithms for four types of noise with different SNRs are shown in Fig. 6. The results show that the ATF-based robust algorithm outperforms the other two algorithms. In fact, the ATF-based robust algorithm achieves higher recognition rate than the TF-based robust algorithm by 3% on average. This shows that the ATF parameter outperforms the TF parameter for word boundary detection in noisy environment. The main reason is that ATF can adaptively extracts useful frequency information based on the mel-scale frequency bank.

By our previous study, there are two factors affecting the distribution of useful frequency bands for the ATF parameter. One is SNR and the other is the characteristics of the noise condition. In our experiments (with a total of 600 utterances), the probability that each band is selected as useful band for contributing to the frequency energy under different noise conditions and SNRs is shown in Table II. We used three noise conditions to see the effect of noise characteristics; they are multitalker babble noise, cockpit noise, and white noise. From the table, we have the following observations. At high SNR (clean, greater than 20 dB), the probability of each band being selected as useful band in the clean environment is nearly the same, since little noise contaminates any band in this case. At medium SNR (20 dB, 15 dB, 10 dB), the useful bands gradually concentrate on the low frequency part (band 1$\sim$band 7). This is because the energy of noise usually concentrates on high frequency bands. This concentration situation is more obvious at low SNR (5 dB, 0 dB), where the low-frequency speech signals are less contaminated. When considering different noise conditions, we find that the useful bands under white noise and cockpit noise with low SNRs more concentrate in low frequency than those under the multitalker babble noise. In other words, the useful bands under the multitalker babble noise spread wider than those under the other two types of noise. The reason is that the characteristics of multitalker babble noise are very similar to those of word signals.

TABLE II
PROBABILITY OF EACH FREQUENCY BAND BEING SELECTED AS A USEFUL BAND FOR CONTRIBUTING TO THE FREQUENCY ENERGY UNDER DIFFERENT NOISE CONDITIONS AND SNRS

White Noise

SNR	Band index																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Clean	5.4	5.6	5.6	5.6	5.5	5.6	5.6	5.5	5.1	4.3	3.9	4.1	4.1	4.2	4.4	5.4	5.3	5.3	.5.1	4.6
20dB	4.5	9.8	9.6	8.9	7.5	7.0	6.4	5.5	4.3	3.8	3.6	3.8	3.3	2.8	3.0	3.5	3.4	3.4	3.2	2.9
15dB	3.1	21.6	18.8	14.7	7.3	7.6	4.9	3.5	2.0	1.3	1.3	1.1	1.0	1.3	1.7	1.6	1.6	2.0	1.8	1.7
10dB	1.6	31.8	27.5	18.0	5.4	6.2	4.1	2.4	1.5	0.4	0.7	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
5dB	5.7	31.5	25.4	15.8	5.6	5.7	4.0	3.1	1.3	0.5	0.8	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0-	0.0
0dB	12.6	30.4	22.8	12.1	5.6	5.7	5.3	3.0	1.0	0.4	0.6	0.2	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0

(a)

Babble Noise

SNR	Band index																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Clean	5.4	5.6	5.6	5.6	5.5	5.6	5.6	5.5	5.1	4.3	3.9	4.1	4.1	4.2	4.4	5.4	5.3	5.3	5.1	4.6
20dB	4.3	8.0	7.6	7.4	6.5	6.3	6.2	5.9	5.1	4.0	3.7	4.4	3.7	2.9	3.0	4.3	4.8	4.2	4.2	3.6
15dB	3.3	12.2	9.4	8.9	7.0	6.3	6.1	5.4	4.3	2.7	3.7	4.3	2.8	1.6	1.9	3.1	4.0	4.2	4.7	4.1
10dB	2.4	11.8	9.1	7.3	4.6	3.5	4.8	5.8	5.0	3.1	4.7	4.4	3.8	2.0	1.4	1.8	3.9	5.9	7.5	7.1
5dB	5.2	7.7	4.9	4.8	3.9	2.2	4.8	5.3	4.8	4.3	5.9	5.7	3.4	3.1	2.3	1.8	4.0	8.4	8.9	8.7
0dB	5.4	5.3	3.7	3.7	3.3	2.2	4.6	5.8	6.4	6.4	7.6	5.9	4.8	3.4	3.2	2.3	3.8	6.7	8.1	7.6

(b)

Cockpit Noise

SNR	Band index																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Clean	5.4	5.6	5.6	5.6	5.5	5.6	5.6	5.5	5.1	4.3	3.9	4.1	4.1	4.2	4.4	5.4	5.3	5.3	5.1	4.6
20dB	4.4	8.6	8.4	8.1	7.2	7.1	6.3	5.4	4.4	3.3	3.7	3.4	3.0	2.8	3.0	4.2	3.3	4.9	4.7	3.8
15dB	3.7	15.7	13.9	12.9	8.6	9.1	7.2	4.0	1.8	1.2	1.7	2.1	1.5	1.3	1.9	2.5	1.3	3.4	3.4	3.0
10dB	4.5	19.3	18.5	15.2	9.6	10.7	8.6	4.1	1.9	0.4	0.6	1.5	0.5	0.1	0.1	0.3	0.1	1.0	2.0	0.9
5dB	8.0	15.4	16.9	14.6	10.0	11.0	10.4	4.6	1.3	0.3	1.0	2.1	0.5	0.2	0.1	0.3	0.0	1.3	1.3	0.6
0dB	11.4	13.0	14.4	13.8	10.5	10.9	10.0	4.5	2.6	0.7	0.9	1.7	0.7	0.2	0.2	0.4	0.1	2.4	1.1	0.6

(c)

They are both pronounced by human, not by the machine or the others. Hence, the energy distribution of the additive (multitalker babble) noise on each band is very close to that of the word signal on each band.

Our experiments show that the ATF parameter can extract useful frequency information (energy) of word signal by using the first N_a maximum total-energy bands without the need to recognize SNRs and the types of noise. It can reliably yield obvious word boundary and can be used to distinguish the word signal from noise. The main feature of the ATF parameter is its ability to track the properties of varying noisy environment. The reliability of the ATF parameter less depends on SNRs and the characteristics of noise due to its adaptation ability.

The robust algorithm developed in [5] and used in the above experiments requires empirically chosen thresholds and ambiguous rules, which are not easily determined by human. In fact, in our experiments, we have made every effort in tuning the thresholds of the robust algorithm to get the best performance for our noisy speech database and test environment. Some researchers used neural network's learning ability to attack this tuning problem. Even though, the number of hidden layer and the number of the nodes per hidden layer for a neural network still need to be determined. The lack of an effective method to learn the network structure is usually a drawback. In the following section, we shall use a neural fuzzy network with structure learning ability to develop another robust word boundary detection scheme based on the proposed ATF parameter.

III. NEURAL FUZZY NETWORK FOR WORD BOUNDARY DETECTION

In this section, we shall first introduce a neural fuzzy network and then propose a robust word boundary detection scheme based on this network with the ATF parameter as input pattern.

A. Self-Constructing Neural Fuzzy Inference Network

The neural fuzzy network that we used for word boundary detection is called the self-constructing neural fuzzy inference network (SONFIN) that we proposed previously in [11]. The SONFIN is a general connectionist model of a fuzzy logic system, which can find its optimal structure and parameters automatically. There are no rules initially in the SONFIN, and they are created and adapted as on-line learning proceeds via simultaneous structure and parameter learning. The SONFIN can always find itself an economic network size, and the learning speed as well as the modeling ability are all superior to normal neural networks.

The structure of the SONFIN is shown in Fig. 7(a). This six-layered network realizes a fuzzy model of the following form:

$$\text{Rule } i\text{: IF } x_1 \text{ is } A_{i1} \text{ and} \cdots \text{and } x_n \text{ is } A_{in}$$
$$\text{THEN } y \text{ is } m_{0i} + a_{ji}x_j + \cdots \tag{18}$$

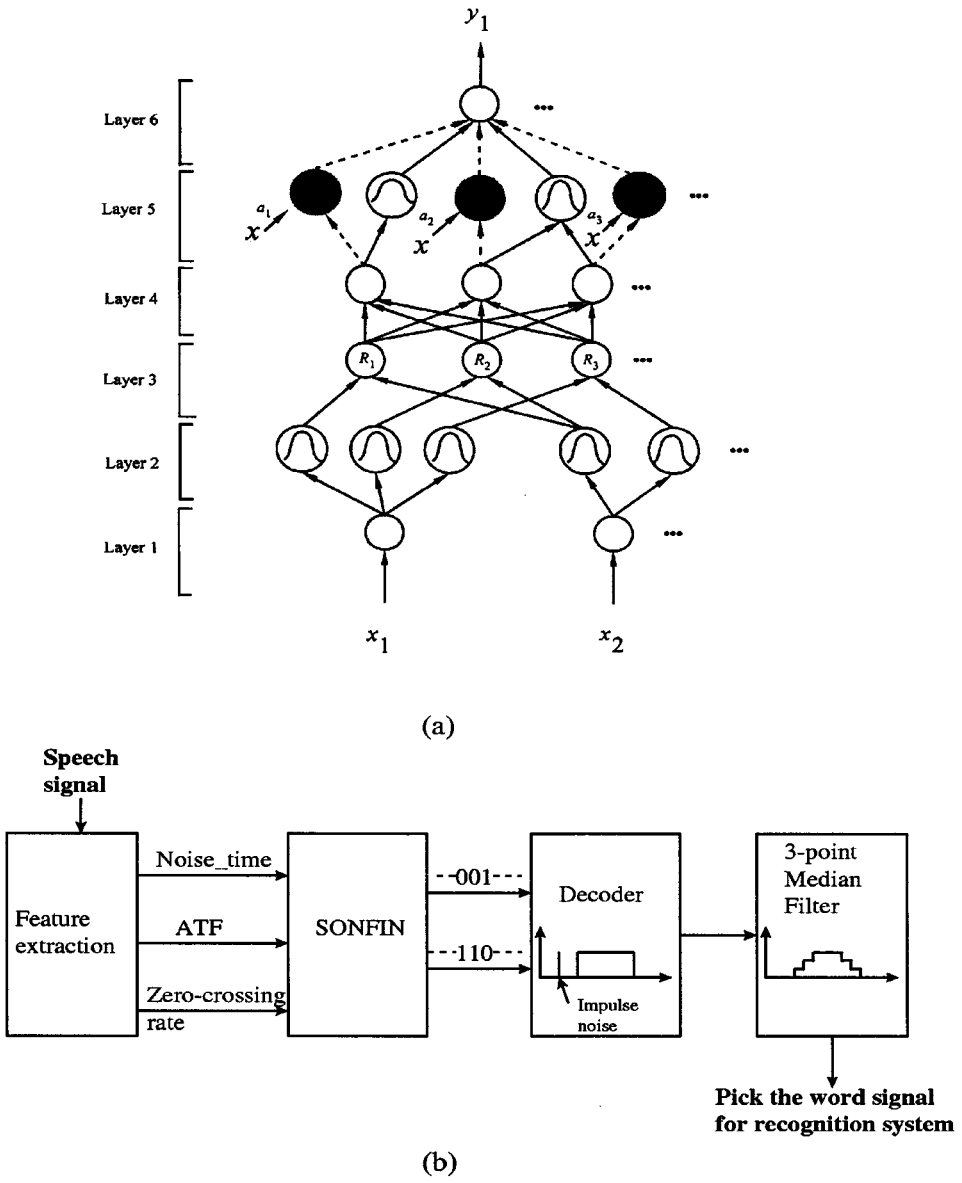


Fig. 7. (a) Network structure of the SONFIN. (b) Flowchart of the SONFIN-based word boundary detection procedure.

where

A_{ij} fuzzy set;

m_{0i} center of a symmetric membership function on y;

a_{ji} consequent parameter.

It is noted that unlike the traditional TSK model [12], [21], [22] where all the input variables are used in the output linear equation, only the significant ones are used in the SONFIN, i.e., some a_{ij}'s in the above fuzzy rules are zero. We shall next describe the functions of the nodes in each of the six layers of the SONFIN.

Each node in Layer 1, which corresponds to one input variable, only transmits input values to the next layer directly. Each node in Layer 2 corresponds to one linguistic label (small, large, etc.) of one of the input variables in Layer 1. In other words, the membership value which specifies the degree to which an input value belongs a fuzzy set is calculated in Layer 2. A node in Layer 3 represents one fuzzy logic rule and performs precondition matching of a rule. The number of nodes in layer 4 is equal to that in Layer 3, and the result (firing strength) calculated in Layer 3 is normalized in this layer. Layer 5 is called the consequent layer. Two types of nodes are used in this layer, and they are denoted as blank and shaded circles in Fig. 7(a), respectively. The node denoted by a blank circle (blank node) is the essential node representing a fuzzy set of the output variable. The shaded node is generated only when necessary. One of the inputs to a shaded node is the output delivered from Layer 4, and the other possible inputs (terms) are the selected significant input variables from Layer 1. Combining these two types of nodes in Layer 5, we obtain the whole function performed by this layer as the linear equation on the THEN part of the fuzzy logic rule in (18). Each node in Layer 6 corresponds to one output variable. The node integrates all the actions recommended by Layer 5 and acts as a defuzzifier to produce the final inferred output.

Two types of learning, structure and parameter learning, are used concurrently for constructing the SONFIN. The structure

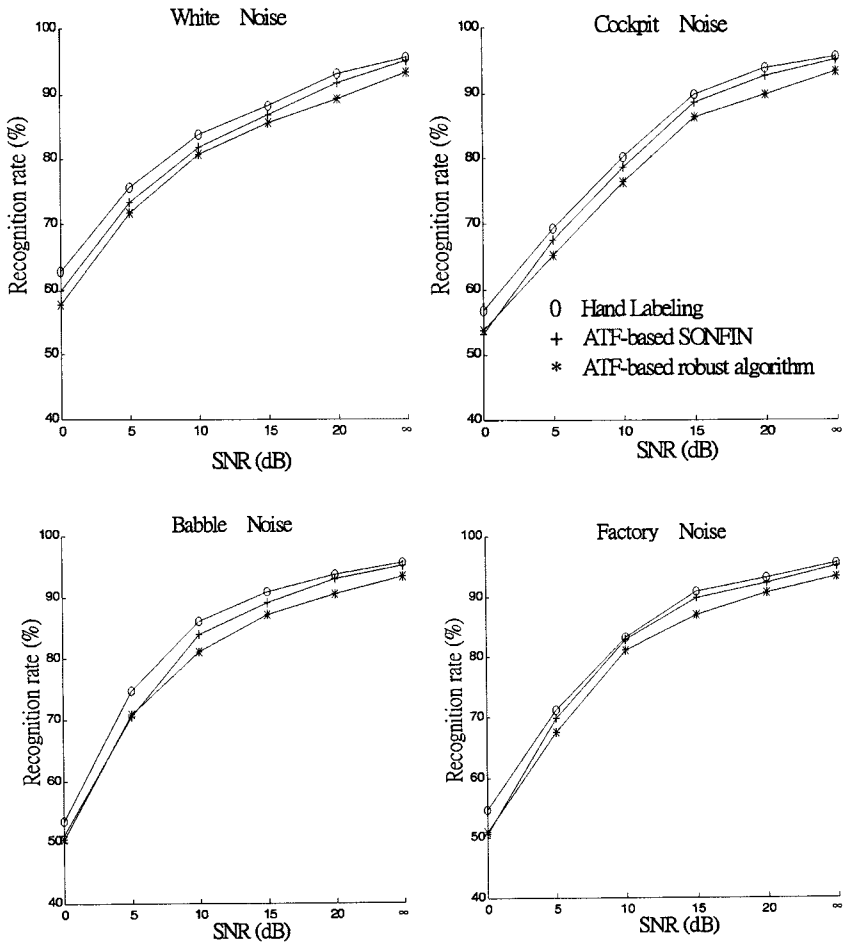


Fig. 8. Recognition rates of three word boundary detection algorithms (ATF-based SONFIN, ATF-based robust algorithm, and hand labeling) in an MTDC-based recognition system across six SNRs and four noise conditions.

learning includes both the precondition and consequent structure identification of a fuzzy if-then rule. For the parameter learning, based upon supervised learning algorithms, the parameters of the linear equations in the consequent parts are adjusted to minimize a given cost function. The SONFIN can be used for normal operation at any time during the learning process without repeated training on the input–output patterns when on-line operation is required. There are no rules in the SONFIN initially, and they are created dynamically as learning proceeds upon receiving on-line incoming training data by performing the following learning processes simultaneously: a) input/output space partitioning, b) construction of fuzzy rules, c) optimal consequent structure identification, and d) parameter identification. Processes a), b), and c) belong to the structure learning phase and process D belongs to the parameter learning phase. The details of these learning processes can be found in [11].

B. SONFIN for Word Boundary Detection

The procedure of using the SONFIN for word boundary detection is illustrated in Fig. 7(b). The input feature vector of the SONFIN is a combination of the average energy of background noise (Noise_time), adaptive time-frequency (ATF) parameter, and zero-crossing rate. The three parameters in an input feature vector are obtained by analyzing a frame of signal. Hence there are three (input) nodes in Layer 1 of the SONFIN. Here the noise energy, Noise_time, as in (17), is the average of the logarithm of the rms energy on the first five frames of "relative silence" at the beginning of the recording. Before entering the SONFIN, the three input parameters are normalized to be in [0, 1]. For each input vector (corresponding to a frame), the output of SONFIN indicates whether the corresponding frame is a word signal or noise. For this purpose, we used two (output) nodes in Layer 6 of the SONFIN, where the output vector of (1, 0) standing for word signal, and (0, 1) for noise.

The SONFIN was trained by a set of 80 training patterns, which were randomly selected from four noise conditions with different SNRs. These training patterns are classified as word signal or noise by using waveform, spectrum displays and audio output. Among the 80 training patterns, 40 patterns are from word sound category with the desired SONFIN output vector being (1, 0), and the other 40 from noise category with the desired SONFIN output vector being (0, 1). We usually used the frames around the word-noise transition area as the training patterns, because these ambiguous training patterns make the SONFIN get more accurate word boundary in noisy environment. After training, there were only 14 rules generated in the SONFIN.

The SONFIN after training is ready for word boundary detection. As shown in Fig. 7(b), the outputs of the SONFIN are processed by a decoder. The decoder decodes the SONFINs output

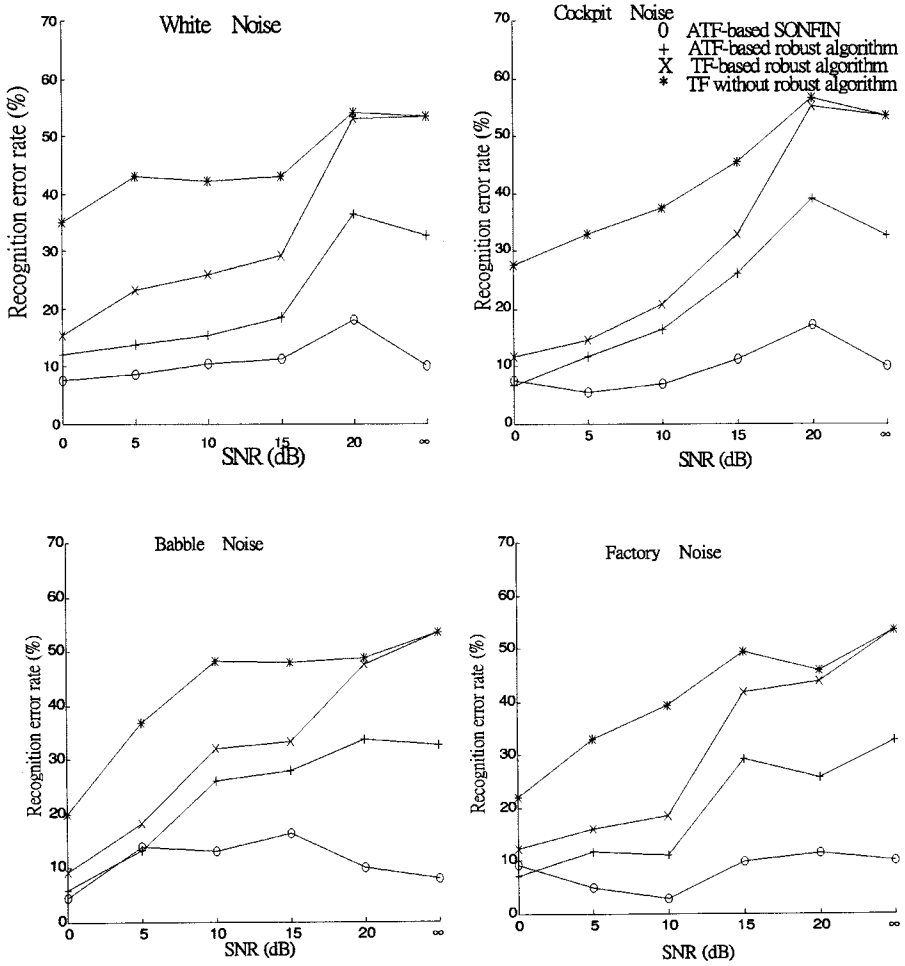


Fig. 9. Recognition error rates of four word boundary detection algorithms (ATF-based SONFIN, ATF-based robust algorithm, TF-based robust algorithm, and TF without robust algorithm) in an MTDC-based recognition system across six SNRs and four noise conditions.

vector (1, 0) as value 100 standing for word signal, and (0, 1) as value 0 standing for noise. In addition, we let the output waveform of the decoder pass through a three-point median filter to eliminate the undesired "impulse" noise. Finally, we recognize the word-signal island as the part of the filtered waveform whose magnitude is greater than 30, and duration is long enough (by setting a threshold value). We then send the part of original signal corresponding to the allocated word-signal island to our word recognition system.

To verify the performance of the SONFIN for word boundary detection, the experiments performed in Section II-C are performed here again, with the robust algorithm replaced by the SONFIN. The results are shown in Fig. 8. Again, the performance evaluation of word boundary detection is based on the recognition rate of the same speech recognition system as in Section II-C. In Fig. 8, we also show the performance of the robust algorithm with ATF parameter used in Section II-C, and the performance of hand labeling (i.e., manually determined boundaries.) Considering another performance index, we examined the recognition error rates averaged across the four noise conditions due to word boundary detection as a function of SNRs as shown in Fig. 9. Here, the recognition error rate is the ratio of the recognition errors due to wrong word boundary detection (taking recognition scores obtained with hand-labels as a reference) to the total number of recognition errors of the detection algorithm [5]. More precisely, let the recognition errors obtained when using hand labeling be $E_{\rm hl}$, and the recognition errors obtained when using automatic word boundary detection algorithm be $E_{\rm al}$. Then the recognition error rate is given by $(E_{\rm al} - E_{\rm hl})/E_{\rm al}$. This index represents the percentage of recognition errors attributable to word boundary detection errors relative to the total number of errors, where the recognition rate with hand-labeled boundaries is used as a reference. These results show that, by using the same three parameters (Noise_time, ATF, and zero-crossing rate), the SONFIN outperforms the robust algorithm by about 2% in recognition rate. As a total, the ATF-based SONFIN had higher recognition rate than the TF-based robust algorithm in [5] by about 5%. Also, the ATF-based SONFIN reduced the recognition error rate due to endpoint detection to about 10%, compared to about 20% obtained with the ATF-based robust algorithm, about 30% obtained with the TF-based robust algorithm, about 40% obtained with the TF without robust algorithm, and about 50% obtained with the modified version of the Lamel *et al.* algorithm [3], [4]. We also found that the SONFIN could approach the result of hand labeling, which is usually considered as the optimum result for reference. Notice that, in the above tests, all the thresholds of the robust algorithm were tuned exhaustively to achieve

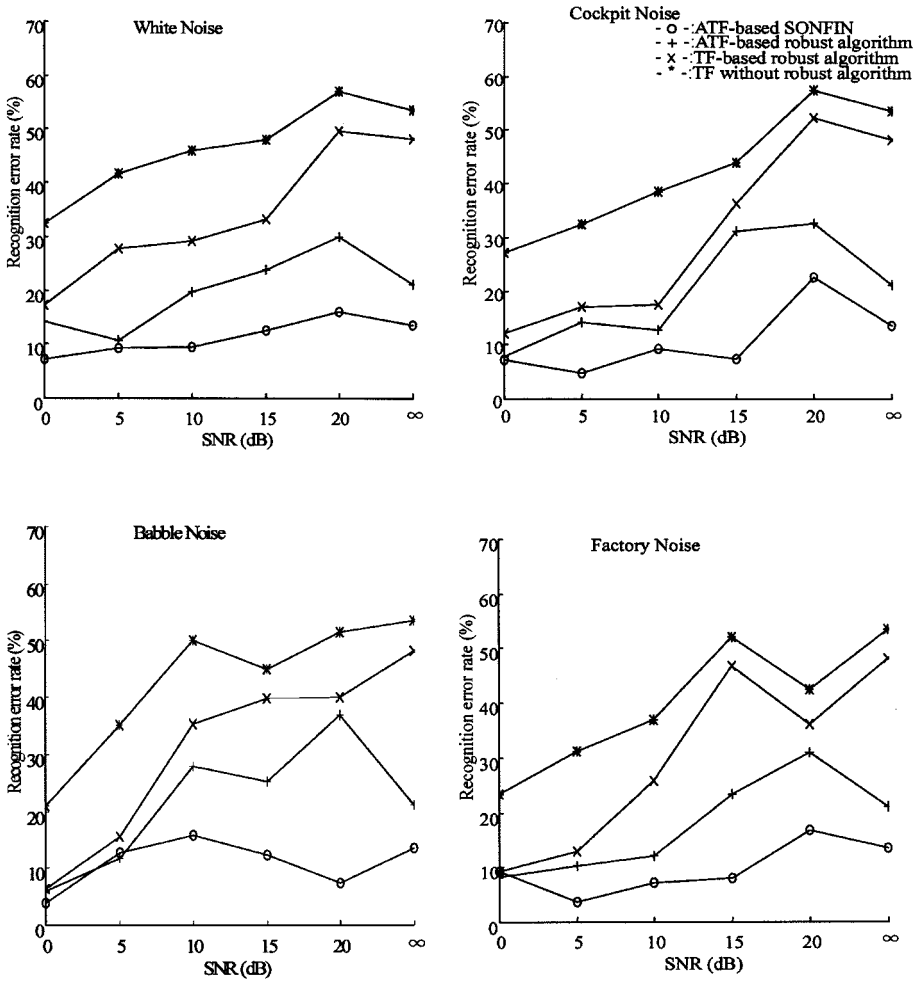


Fig. 10. Recognition error rates of four word boundary detection algorithms (ATF-based SONFIN, ATF-based robust algorithm, TF-based robust algorithm, and TF without robust algorithm) in an MFCC-based HMM recognition system across six SNRs and four noise conditions.

the best performance in our test environments. In the hand labeling process for word boundary detection, we allowed three trials. That is, if the hand-labeled segment of the test signal was not recognized correctly by the speech recognizer, the signal was relabeled again manually. If the second trial failed, we performed the third trial. A hand-labeled segment was classified as an error word boundary only when all the three trials failed. This rigor hand labeling process might be the reason that the average recognition error rate of the TF-based robust algorithm in our tests is higher than that reported in [5].

After learning, the SONFIN generated ten membership functions in the input dimension (variable) "Noise_time" representing the energy of environment noise [see (17)]. In other words, the SONFIN automatically classified the energy of background noise into ten *fuzzy* categories. As compared to the robust algorithm in [5] (see Section II-C) which classified the noise energy into three *crispy* levels (high, medium, and low) by empirically human determination, the SONFIN provided more precise noise classification by self-learning. Similarly, the SONFIN automatically classified the other two features, ATF and zero-crossing rate, into 11 and 14 *fuzzy* categories by learning proper membership functions. This is the important reason why the SONFIN could get reasonably high classification rate by using only 14 rules for the word boundary detection in the noisy environment. Due to the self-learning ability of SONFIN, the proposed algorithm avoids the need of empirically determining thresholds and ambiguous rules in normal word boundary detection algorithms. This is the first motivation for us to use SONFIN in such an application. Due to the physical meaning of fuzzy if-then rule, each input node in the SONFIN is only connected to its related rule nodes through its term nodes, instead of being connected to all the rule nodes in Layer 3 of the SONFIN. This results in a small number of weights to be tuned in the SONFIN. In contrast, in the normal fully connected neural networks such as backpropagation network and radical basis function network [23], the number of tuning weights is usually large as compared to the number of rule nodes learned in the SONFIN for the same learning task. This forms our second motivation for the use of SONFIN in this paper.

In order to see the performance of our algorithms on other speech features and recognizer, we replace the MTDC-based recognizer used in the previous experiments by the MFCC (mel-frequency cepstral coefficient)-based HMM recognizer with temporal filter in another set of experiments, where the temporal filter is used to remove the noise components in the feature extraction phase. The number of coefficients of each frame used in this HMM recognizer is 26, including MFCCs,

energy, delta MFCCs and delta energy, and the analysis order is 24. Each Mandarin digits is modeled by a five-state, left-to-right, continuous density HMM. In the HMM, each state is split into two streams, and a mixture Gaussian density with two mixture components in each stream is assigned to each state observation probability. The recognition error rates averaged across the four noise conditions due to word boundary detection as a function of SNRs are shown in Fig. 10. The results show that the conclusions on the good performance of the proposed word boundary detection algorithms still hold on the common speech features and recognizer.

Although the SONFIN has the advantages of small network size, high learning speed, and high learning accuracy, its merits are obtained at the expense of longer CPU time. The CPU time of these algorithms running in Pentium 90 for processing 100 frames of a speech signal is recorded as follows:

- Feature extraction procedures:
  TF parameter (0.4945 s),
  ATF parameter (0.5494 s).
- Word detection algorithms:
  TF-based robust algorithm (0.5011 s),
  ATF-based SONFIN (0.9513 s).

Extracting the ATF parameter needs more computation time than the TF parameter due to the multiband analysis of the former as compared to the single-band analysis of the latter. The computation time of SONFIN is mainly taken in calculating the Gaussian membership functions, compared to the crispy decision rules (threshold comparisons) in the robust algorithm.

IV. Conclusions

In this paper, we have proposed a reliable parameter, *adaptive time-frequency* (ATF), that possesses both the time and frequency features for word boundary detection in noisy environment. This parameter adaptively adopts some useful bands from 20 mel-scale frequency bands for producing useful frequency feature to enhance time feature in noisy environment. Comparative study has shown that the ATF parameter is very beneficial for several SNRs and noise conditions (including clean speech, for which very good results were obtained.) The new ATF parameter increased by about 3% the recognition rate of a robust word detection algorithm, which adopts the original TF parameter (TF-based robust algorithm). It also reduced the recognition error rate due to endpoint detection to about 20%, compared to an average of approximately 30% obtained with the TF-based robust algorithm, 40% obtained with TF without robust algorithm, and 50% obtained with the modified version of the Lamel *et al.* algorithm. Based on the ATF parameter, we have also proposed a word boundary detection scheme based on a neural fuzzy network, SONFIN. The SONFIN can learn by itself the (fuzzy) word boundary detection rules and the classification of background noise. The performance of the SONFIN with ATF parameter has been evaluated across several SNRs and noise conditions. Our experiments showed that the proposed scheme (ATF-based SONFIN) achieved higher recognition rate by about 2% than the ATF-based robust algorithm, and thus by about 5% than the TF-based robust algorithm. On the other performance index, the ATF-based SONFIN reduced the recognition error rate due to endpoint detection to about 10%, compared to about 20% obtained with the ATF-based robust algorithm.

References

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297–315, Feb. 1975.

[2] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, vol. 8, pp. 45–60, 1989.

[3] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "An improved endpoint detector for isolated word recognition," *IEEE ASSP Mag.*, vol. 29, pp. 777–785, Aug. 1981.

[4] B. Reaves, "Comments on an improved endpoint detector for isolated word recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 526–527, Feb. 1991.

[5] J. C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 406–412, July 1994.

[6] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier," *IEEE Tran. Speech Audio Processing*, vol. 1, pp. 250–255, Apr. 1993.

[7] S. J. Kia and G. G. Coghill, "A mapping neural network and its application to voiced-unvoiced-silence classification," in *Proc. 1st New Zealand Int. Two-Stream Conf. Artificial Neural Networks Expert Systems*, 1993, pp. 104–108.

[8] M. Hamada, Y. Takizawa, and T. Norimatsu, "A noise robust speech recognition," in *Proc. Int. Conf. Speech Language Processing '90*, 1990, pp. 893–896.

[9] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[10] T. Ghiselli-Crippa and A. El-Jaroudi, "A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech," *Proc. Int. Conf. Speech Language Processing '91*, vol. 1, pp. 441–444, 1991.

[11] C. F. Juang and C. T. Lin, "An on-line self-constructing neural fuzzy inference network and its application," *IEEE Trans. Fuzzy Syst.*, vol. 6, pp. 12–32, Feb. 1998.

[12] C. T. Lin and C. S. G. Lee, *Neural Fuzzy Systems: A Neural-Fuzzy Synergism to Intelligent Systems*. Englewood Cliffs, NJ: Prentice-Hall, May 1996.

[13] C. T. Lin, *Neural Fuzzy Control Systems with Structure andParameter Learning*. Singapore: World Scientific, 1994.

[14] J. B. Allen, "Cochlear modeling," *IEEE Acoust., Speech, Signal Processing Mag.*, vol. 2, pp. 3–29, 1985.

[15] D. O'Shaughnessy, *Speech Communication*. Reading, MA: Addision-Wesley, 1987, p. 150.

[16] Y. Ariki, S. Mizuta, and T. Sakai, "Spoken-word recognition using dynamic features analyzed by two-dimensional cepstrum," *Proc. Inst. Elect. Eng.*, vol. 136, Apr. 1989.

[17] H. F. Pai and H. C. Wang, "A study on two-dimensional cepstrum approach for speech recognition," *Comput. Speech Lang.*, vol. 6, pp. 361–375, 1992.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

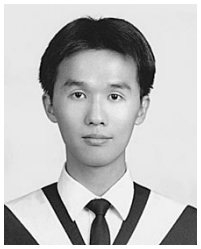
[19] C. T. Lin, H. W. Nein, and J. Y. Hwu, "GA-based noisy speech recognition using two-dimensional cepstrum," IEEE Trans. Speech Audio Processing, to be published.

[20] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[21] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, pp. 116–132, Jan. 1985.

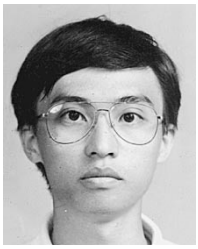
[22] M. Sugeno and K. Tanaka, "Successive identification of a fuzzy model and its applications to prediction of a complex system," *Fuzzy Sets Syst.*, vol. 42, no. 3, pp. 315–334, 1991.

[23] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–579, July 1993.



Gin-Der Wu received the B.S. degree in engineering science from National Cheng-Kung University, Taiwan, R.O.C., in 1996. He is currently pursuing the Ph.D. degree in electrical and control engineering from the National Chiao-Tung University, Hsinchu, Taiwan.

His current research interests are speech recognition and enhancement in noisy environment, adaptive signal processing, neural networks, and fuzzy control.



Chin-Teng Lin (M'91–SM'99) received the B.S. degree in control engineering from National Chiao-Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 1986 and the M.S.E.E. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, in 1989 and 1992, respectively.

Since August 1992, he has been with the College of Electrical Engineering and Computer Science, NCTU, where he is currently a Professor of electrical and control engineering. He has also been Deputy Dean of the Research and Development Office, NCTU, since 1998. His current research interests are fuzzy systems, neural networks, intelligent control, human-machine interface, and video and audio processing. He is the coauthor of *Neural Fuzzy Systems—A Neuro-Fuzzy Synergism to Intelligent Systems* (Englewood Cliffs, NJ: Prentice-Hall), and the author of *Neural Fuzzy Control Systems with Structure and Parameter Learning* (Singapore: World Scientific). He has published more than 50 journal papers in the areas of neural networks and fuzzy systems.

Dr. Lin is a member of Tau Beta Pi and Eta Kappa Nu. He is also a Member of the IEEE Computer Society, the IEEE Robotics and Automation Society, and the IEEE Systems, Man, and Cybernetics Society. He has been the Executive Council Member of the Chinese Fuzzy System Association (CFSA) since 1995, and the Supervisor of Chinese Automation Association since 1998. He was the Vice Chairman of IEEE Robotics and Automation Taipei Chapter in 1996 and 1997. He won the Outstanding Research Award granted by National Science Council (NSC), Taiwan, for 1997–1998 and 1999–2000, the Outstanding Electrical Engineering Professor Award granted by the Chinese Institute of Electrical Engineering (CIEE) in 1997, and the Outstanding Engineering Professor Award granted by the Chinese Institute of Engineering in 2000.